Title:

Subtitle

May 15, 2024

Abstract

1 Introduction

The National Institute of Standards and Technology Artificial Intelligencde (AI) Risk Management Framework (RMF).[1]

2 Generative AI Governance

3 Generative AI Inventories

4 Generative AI Risk Tiers

5 Generative AI Risk Measurement

6 Generative AI Risk Management

Conclusion

Acknowledgments

Thank you to Bernie Siskin and Nick Schmidt of BLDS and Eric Sublett of Relman Colfax for formative discussions relating to GAI risk tiering.

Abbreviations

- AI: Artificial Intelligence
- AI RMF: Artificial Intelligence Risk Management Framework
- GAI: Generative AI
- RMF: Risk Management Framework

[1] NIST AI. Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023.

Appendix A: Generative AI–Trustworthy Characteristic Crosswalk

Appendix B: Example Risk Tiers for Generative AI

Appendix C: List of Publicly Available Model Testing Suites (“Evals”)

C.1: Publicly Available Model Testing Suites (“Evals”) by Trustworthy Characteristic

C.2: Publicly Available Model Testing Suites (“Evals”) by Generative AI Risk

Appendix D: List of Common Adversarial Prompting Strategies

D.1: Common Adversarial Prompting Strategies by Trustworthy Characteristic

D.2: Common Adversarial Prompting Strategies by Generative AI Risk

Appendix E: Common Risk Controls for Generative AI

E.1: Common Risk Controls for Generative AI by Trustworthy Characteristic

E.2: Common Risk Controls for Generative AI by Generative AI Risk

Appendix F: Example Low-risk Generative AI Measurement and Management Plan

6.1 F.1: Example Low-risk Generative AI Measurement and Management Plan by Trustworthy Characteristic

6.2 F.2: Example Low-risk Generative AI Measurement and Management Plan by Generative AI Risk

Appendix G: Example Medium-risk Generative AI Measurement and Management Plan

6.3 G.1: Example Medium-risk Generative AI Measurement and Management Plan by Trustworthy Characteristic

6.4 G.2: Example Medium-risk Generative AI Measurement and Management Plan by Generative AI Risk

Appendix H: Example High-risk Generative AI Measurement and Management Plan

6.5 H.1: Example High-risk Generative AI Measurement and Management Plan by Trustworthy Characteristic

6.6 H.2: Example High-risk Generative AI Measurement and Management Plan by Generative AI Risk